

MASARYKOVA UNIVERZITA
INNOVATION LECTURES (I.N.N.O.I.E.C.) www.muni.cz

Binding and Kinetics for Experimental Biologists
Lecture 7
Dealing with uncertainty: Confidence intervals

Petr Kuzmič, Ph.D.
BioKin, Ltd.
WATERTOWN, MASSACHUSETTS, U.S.A.

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

EVROPSKÁ UNIE **esf** MINISTERSTVO ŠKOLSTVÍ, Mládeže a Tělovýchovy OP Vzdělávání pro konkurenceschopnost UNIVERZITA MASARYKOVA BRNO

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

BioKin I.L.B.


Part 1: Confidence intervals by systematic searching

"Profile-t" method


BKEB Lec 7: Confidence Intervals 4

"Hunches and intuitive impressions are essential for getting the work started, but it is only through the **quality of the numbers** at the end that the **truth** can be told."
-Lewis Thomas

L. Thomas (1977) "Biostatistics in Medicine", Science 198, 675



Gregor Mendel (1822-1884)



approximately 3:1 G / Y

Google - July 20, 2011

But how much **confidence** can you have in that number?

BKEB Lec 7: Confidence Intervals 2

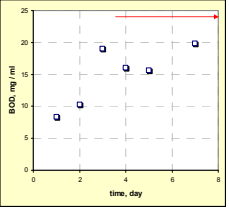
Example problem: Biological oxygen demand (B.O.D.)

A CLASSIC DATA SET IN STATISTICAL LITERATURE

BOD = measure of organic pollution in environmental water

Table A1.4 Biochemical oxygen demand versus time.

Time (days)	Biochemical Oxygen Demand (mg/l)	Time (days)	Biochemical Oxygen Demand (mg/l)
1	8.3	4	16.0
2	10.3	5	15.6
3	19.0	7	19.8



BOD at $t \rightarrow \infty$?

Bates D. M. & Watts, D. G. (1988) *Nonlinear Regression and its Applications* Wiley, New York, p. 270

BKEB Lec 7: Confidence Intervals 5

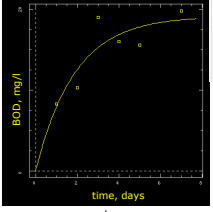
Lecture outline

- The problem:**
How much (or how little) can we trust our rate and equilibrium constants?
- The solution:**
Always report at least *some* measure of parameter uncertainty:
 - formal standard error
 - confidence interval
 - (a) by systematic search (profile-t method)
 - (b) by stochastic simulations (Monte-Carlo method)
- An implementation:**
Software *DynaFit*.
- An example:**
The classic "biological oxygen demand (BOD)" problem

BKEB Lec 7: Confidence Intervals 3

Theoretical model: Exponential growth

COMPARE ALGEBRAIC MODEL WITH DYNAFIT NOTATION



Optimized Parameters

No.	Para/Set	Initial	Final
#1	k	1	0.531091
#2	[Oxygen]	10	19.1426

BOD_{max} = 19.1 mg/l

ALGEBRAIC MODEL:
 $B = B_{\max} [1 - \exp(-kt)]$

DYNAFIT MODEL:
[mechanism]
Oxygen \rightarrow Bacteria : k

BKEB Lec 7: Confidence Intervals 6

How much should we trust these model parameters?

FIX B_{max} AT AN ARBITRARY VALUE, OPTIMIZE k

optimized parameter
fixed parameter

$B_{max} = 19.1$ $k = 0.53$ sum of squares = 26.0	$B_{max} = 25.0$ $k = 0.28$ sum of squares = 41.9	$B_{max} = 30.0$ $k = 0.20$ sum of squares = 57.6
---	---	---

BioKin
BKEB Lec 7: Confidence Intervals 7

Confidence intervals: Profile-t method in DynaFit

A SEQUENCE OF SEVERAL INDEPENDENT LEAST-SQUARES FITS

INPUT:

```
[mechanism]
Oxygen --> Bacteria : k

[constants]
k = 1 ?

[concentrations]
Oxygen = 10 (?)
```

ALGORITHM:

1. Perform an **initial fit** with all parameters optimized
2. Perform a series of **follow-up fits** focusing on a given parameter
 - 2a. "Freeze" the parameter at values progressively further away from optimal
 - 2b. Optimize all remaining parameters
 - 2c. Repeat (2a) and (2b) until sum of squares reaches a "critical value" above minimum

REFERENCE: Bates, D. M., and Watts, D. G. (1988) *Nonlinear Regression Analysis and its Applications* Wiley, New York, pp. 127-130

BioKin
BKEB Lec 7: Confidence Intervals 10

A little better than nothing: Formal standard errors

THIS IS WHAT MOST PAPERS REPORT IN THE LITERATURE

No.	Par#Set	Initial	Final	Std. Error
#1	k	1	0.531091	0.203092
#2	[Oxygen]	10	19.1426	2.49592

$BOD_{max} = (19.1 \pm 2.5) \text{ mg/l}$

implies the interval

$19.1 - 2.5 = 16.6$
 $19.1 + 2.5 = 21.6$

[settings]
[Output]
InferenceBands = y

BioKin
BKEB Lec 7: Confidence Intervals 8

Confidence level (%) and the width of confidence intervals

HIGHER CONFIDENCE LEVEL = WIDER CONFIDENCE INTERVAL

Upper limit for BOD_{max} could not be determined at 99% confidence level.

BioKin
BKEB Lec 7: Confidence Intervals 11

The correct way to do it: Approximate confidence intervals

VERY RARELY REPORTED IN THE LITERATURE (UNFORTUNATELY)

No.	Par#Set	Initial	Final	Std. Error	CV (%)	Low	Low P (%)	High	High P (%)
#1	k	1	0.531091	0.203092	38.24	0.203097	90	1.2564	90
#2	[Oxygen]	10	19.1426	2.49592	13.04	15.0333	90	29.2602	90

$BOD_{max} = (19.1 \pm 2.5) [15.0 - 29.3] \text{ mg/l}$

BioKin
BKEB Lec 7: Confidence Intervals 9

Example of a half-open confidence interval

UPPER LIMITS FOR BIMOLECULAR ASSOCIATION RATE CONSTANTS OFTEN CANNOT BE DETERMINED

MECHANISM

$$E + S \xrightleftharpoons[k_2]{k_1} ES \xrightarrow{k_3} E + P$$

$$E + I \xrightleftharpoons[k_7]{k_5} EI \xrightleftharpoons[k_6]{k_4} EI^*$$

$k_4 = (5 \pm 200) [3 - \infty] \mu\text{M}^{-1}\text{s}^{-1}$

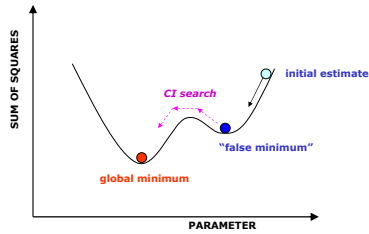
CONFIDENCE INTERVAL FOR k_4

Moss, Kuzmic, et al. (1996) *Biochemistry* 35, 3457-3464.

BioKin
BKEB Lec 7: Confidence Intervals 12

Search for confidence intervals may diagnose “false minima”

AN OCCASIONAL SIDE-BENEFIT OF CONFIDENCE INTERVAL SEARCHES



Monte-Carlo confidence intervals: Algorithm

1. Perform an **initial fit** as usual
2. Perform a **large series** (> 1000) of **follow-up fits**
 - 2a. **Simulate** an artificial data set with **random errors** superimposed in ideal data
 - 2b. Perform a fit of the artificial data
 - 2c. Compile a **histogram** of distribution for model parameters from many repeated fits
 - 2d. Determine the range of plausible values for model parameters from the histograms

REFERENCE:

Straume, M., and Johnson, M. L. (1992)

"Monte-Carlo method for determining complete confidence probability distributions of estimated model parameters"

Methods Enzymol. **210**, 117–129.

SUMMARY: Confidence intervals via *profile-t* method

- Confidence intervals are **asymmetrical** for all *nonlinear* parameters
- Frequently much **wider** (more realistic) than \pm formal standard errors
- Sometimes **half-open** intervals: "better than nothing", e.g. for bimolecular association
- Can have **mechanistic** implications (reversible / irreversible steps)
- Sometimes CI search helps in falling out of **false minima**
- In DynaFit scripts, CIs are requested by the "??" syntax
- Should always be reported with their corresponding **confidence levels** (%)
- CIs are wider at higher confidence levels
- Frequently used confidence levels: **90%, 95%, or 99%**
- Computation can be time consuming (many repeated least-squares fits)

```
[settings]
(Marquardt)
ConfidenceLevel = 90
```

Monte-Carlo confidence intervals: DynaFit input

A SINGLE LINE ADDED TO THE DYNAFIT SCRIPT

```
[task]
task = fit
data = progress
confidence = monte-carlo

[mechanism]
Oxygen ----> Bacteria : k

[constants]
k = 1 ??

[concentrations]
Oxygen = 10 ??
...
```

```
[settings]
(MonteCarlo)
Runs = 1000
ConcentrationErrorPercent = 0
plus a number of other advanced control parameters
```

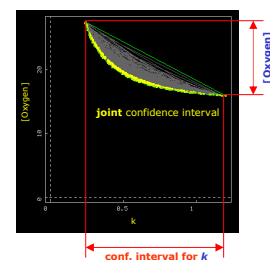
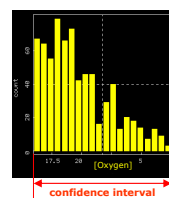
Part 2: Confidence intervals by stochastic simulations

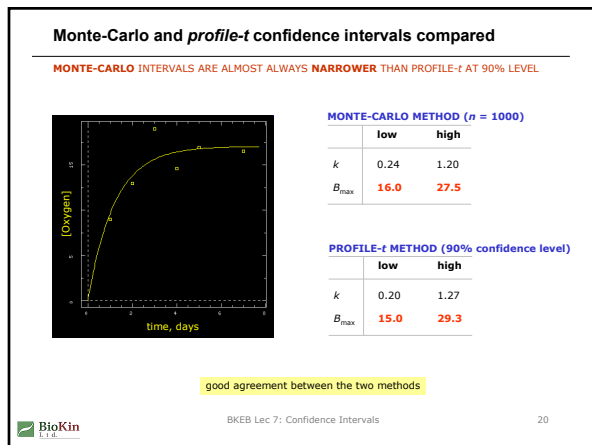
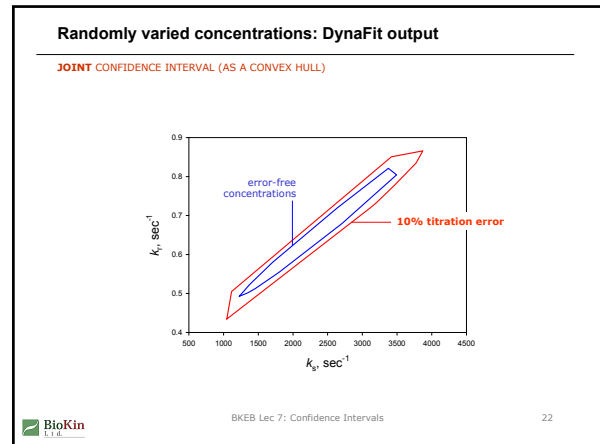
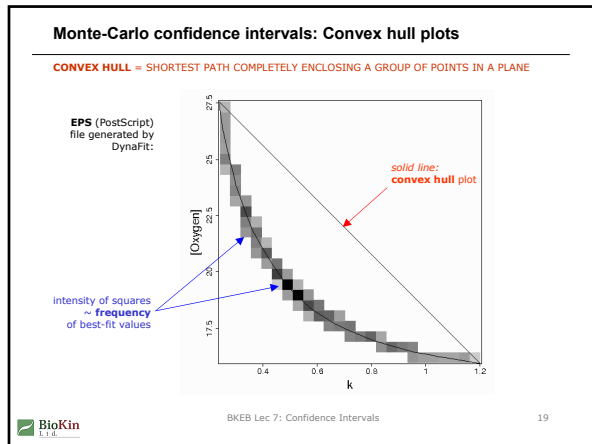
Monte-Carlo method

Monte-Carlo confidence intervals: DynaFit output

HISTOGRAMS OF DISTRIBUTION PLUS CORRELATION PLOTS

Distribution of best-fit values from 1000 least-squares fits of simulated data





SUMMARY: Confidence intervals via Monte-Carlo method

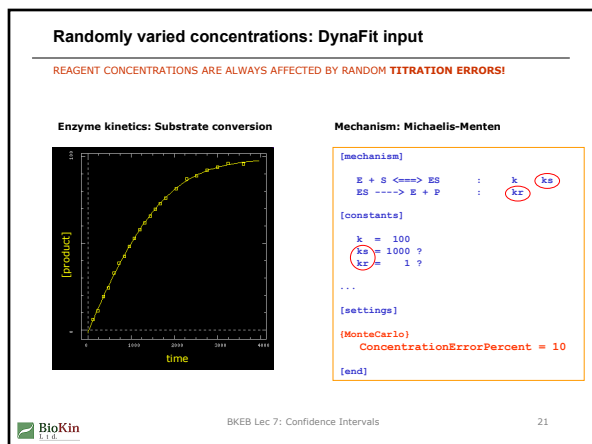
PROS:

- Method makes **no assumptions** about the statistical **distribution** of model parameter errors
- Often uncovers "strange" effects such as **half-open** confidence intervals
 - mechanistic implications (**reversible / irreversible** steps)
- Reveals special patterns in the statistical **correlation** between model parameters
- Does not require an arbitrary choice of **confidence levels (%)**

CONS:

- Method makes **heavy assumptions** about the statistical **distribution** of **experimental errors**
 - could be overcome by the "shuffling" and "shifting" methods in DynaFit
- Can take a **very long time** to compute (multiple hours)
- Does **not** help in discovering false minima

BKEB Lec 7: Confidence Intervals 23



Side comment: The issue of significant digits

BKEB Lec 7: Confidence Intervals 24

Example of poor reporting: Hyperbolic fit in a student project

RESULTS FROM A SEMESTER-LONG RESEARCH PROJECT

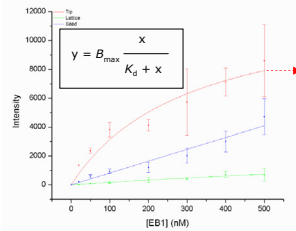


Table 1: K_d values assigned to EB1, EB2 and EB3 at each microtubule configuration

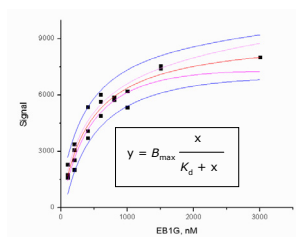
	Tip K_d	Lattice K_d	Seed K_d
EB1	307.09	N/D	N/D

what is wrong with this result?

1. no measure of uncertainty
2. too many digits

Software programs usually report too many digits

OUTPUT GENERATED BY SOFTWARE PACKAGE "ORIGIN"



Parameters	Value	Standard Error
Signal p1	9184.95075	542.54916
Signal p2	642.3346	67.39583

DIRECT OUTPUT FROM SOFTWARE:

$$K_d = (442.3346 \pm 67.39583) \text{ nM}$$

what is wrong with this result?

SENSIBLE WAY TO REPORT IT:

$$K_d = (440 \pm 70) \text{ nM}$$

RECIPE:

1. Round standard error to a **single significant digit**
2. Round best-fit value to the **same number of decimal points**

Overall summary and conclusions

ANY NUMERICAL RESULT REPORTED WITHOUT SOME MEASURE OF UNCERTAINTY IS **MEANINGLESS**

1. Always report at least some measure of statistical uncertainty for all nonlinear model parameters (rate and equilibrium constants).
2. At the very least report the formal \pm standard errors.
3. Confidence intervals are more informative than standard errors.
4. DynaFit offers two different methods for confidence intervals:
 - a. Systematic search (profile- t method)
 - b. Stochastic simulation (Monte-Carlo method)
5. The two methods have their own merits and drawbacks
When in doubt, use both.
6. DynaFit is not a "silver bullet": You must still **use your brain** a lot.